# Supporting Information: Network Analysis of a Proposed Exit Pathway for Protons to the P-side of Cytochrome c Oxidase

Xiuhong Cai[1,2], Kamran Haider[1], Jianxun Lu[1], Slaven Radic[1], Chang Yun Son[3,4], Qiang Cui[3,5], M. R. Gunner[1,2,*]

1. Department of Physics, City College of New York, City University of New York, 160 Convent Avenue New York, NY, 10031, USA
2. Department of Physics, Graduate Center, City University of New York, 365 Fifth Avenue, New York, NY 10016, USA
3. Department of Chemistry and Theoretical Chemistry Institute, University of Wisconsin-Madison, 1101 University Avenue, Madison, WI 53706, USA
4. Current Address: Division of Chemistry and Chemical Engineering, California Institute of Technology, Pasadena, CA, 91125, USA
5. Current Address: Department of Chemistry & Department of Biomedical Engineering, Boston University, 590 Commonwealth Avenue, Boston, MA, 02215, USA

* To whom correspondence should be addressed: mgunner@ccny.cuny.edu.

## Contents of Appendix

## S1. Details of MD simulations

Initial protein coordinates are obtained from the crystal structure of *Rb. sphaeroides* CcO (PDB code 1M56) [1]. MD simulations are carried out in the $P_R$ redox state, with oxidized $Cu_A$, heme *a*, Ferryl heme $a_3$ and Cu(II)-$OH^-$. Titratable residues are modeled in the protonation states suggested by previous MCCE calculations [2]. Glu286, Lys362, Asp407, Lys442 of subunit I, Glu90, Glu185 and Asp251 of subunit III are neutral while all other Glu, Asp, Arg and Lys residues are ionized. His67, Tyr288, His534 of subunit I, Cys252 and Cys256 of subunit II and His37, His132 and His188 of subunit III are ionized and all the other His, Tyr and Cys are neutral.

The MD simulation protocols are carried out as described in reference [3,4], except that the current simulations are at a lower temperature of 303 K and use POPC/POPE instead of DPPC/POPE. All four subunits of the protein are embedded in pre-equilibrated POPC membrane. Water and lipid molecules from the crystal structure are retained, and 9 waters are added to the core cavity region (5 waters) and near Glu286 (4 waters) based on the Grand Canonical Monte Carlo simulation described in reference [5]. The system is solvated with water and KCl ions to keep the system charge neutral and at physiological ion concentration. The resulting system contains 17,700 protein atoms, 302 POPC molecules, 6 POPE molecules, 116 $K^+$ and 107 $Cl^-$ and ≈40,000 water molecules in a 114x114x 138 Å rectangular box. Waters are modeled with TIP3P parameters, the enzyme is modeled using CHARMM22 all-atom force field [6,7] with CMAP correction [7] and the lipids are modeled using the CHARMM36 force field [8]. The atomic partial charge parameters for the cofactors are taken from Johansson *et al* [9]. GROMACS version 4.6.5 [10] was used to perform the simulations. The electrostatic interactions are calculated with the particle mesh Ewald (PME) method [11], and the van der Waals interactions are switched to zero between 10 Å and 12 Å. All bonds involving hydrogen are constrained with the LINCS algorithm [12]. The Nosé-Hoover [13,14] thermostat and Parrinello-Rahman [15] barostat are used to keep the system temperature and pressure at 303 K and 1 atm, respectively.

Trajectories are collected in three protonation substates of the CcO $P_R$ state, called $EP_{a3}^-$, $E^-P_{a3}$ and $EP_{a3}$. The name provides the ionization of E286 and the propionate A of heme $a_3$ ($PRA_{a3}$) (Table S1). The negative superscript indicates the residue is deprotonated and no superscript indicates the group is protonated and thus neutral. Thus, the $EP_{a3}^-$ substate has all four propionic acids deprotonated and Glu286 protonated. In previous simulations by Lu *et al* [4], the summed proton binding of the residues in the Protein Loading Site (PLS) on the P-side of the protein, decreases moving from $E^-P_{a3}$ to $EP_{a3}^-$ to $EP_{a3}$ substates (designated in Lu et al as F1, F2 and F4 ). All four propionic acids are ionized in the $EP_{a3}^-$ MD trajectory, while the $PRA_{a3}$ is protonated in the $E^-P_{a3}$ and $EP_{a3}$. Protonation of $PRA_{a3}$ is strongly correlated with an open the Glu286 cavity.

Each MD system is energy minimized and equilibrated for 1 ns, then subjected to 50 ns of unrestrained MD simulation. The overall protein structure remains stable during the entire simulation, keeping the RMSD value of the entire protein less than 2 Å (Figure S6). The outer copper in $Cu_A$ moves somewhat outward relative to its position found in the crystal structure, but this movement is found to have no impact on the hydrogen bond network identified here.

## S2. Details of MCCE simulations

The MCCE calculations and subsequent hydrogen bond network analysis is performed on the crystal structures of *Rb. sphaeroides* CcO (PDB code: 1M56 [1] and 2GSM [16]) as well as on eleven snapshots extracted at different time points (0ns, 3.5ns, 7ns, 10.5ns, 14ns, 20ns, 26ns, 32ns, 38ns, 44ns and 50ns) from each of the three MD trajectories.

MCCE calculations are carried out as described previously [3,4,17]. For simplicity, only subunits I and II (the core polypeptides) are retained for MCCE analysis. All solvent exposed water molecules and ions with >5% solvent accessible area in the input structure are deleted and replaced with continuum solvent [17]. Buried waters are retained. These can move out of the protein into solution as MCCE incorporates Grand Canonical Monte Carlo sampling for waters and ions [18,19]. The number of waters in the crystal structure and MD snapshots before and after MC sampling are reported in Table S2. GCMC retains 64±2% of the MD waters, perhaps because there is a larger effective van der Waals repulsion in the MCCE force field [20]. The protein dielectric constant is 4. A 35 Å slab of material with a dielectric constant of 4 is added with IPECE [19] to exclude water (with $\varepsilon$=80) around the portion of the protein buried in the membrane. Heme *a*, $a_3$, $Cu_A$ and $Cu_B$ are each combined with their amino acids ligand side chains to form a unified cofactor in MCCE [2,21]. $Ca^{2+}$ and $Mg^{2+}$ are retained as ions buried in protein. The partial charge distribution for the ferryl heme $a_3$ is provided in reference [4].

In MCCE calculations, the residue backbone is fixed but the side chains can have various pre-defined degrees of freedom [17]. Here conformers of residues and ligands sample isosteric degrees of freedom. This allows the hydrogen bond networks to be explored while the C, N and O are retained in positions close to that found in the input MD snapshot or PDB structure. Thus, hydroxyl protons can sit in each torsion minima and in positions that can hydrogen bond to all groups in the neighborhood, both tautomers of neutral His are allowed and Asn and Gln isosteric rotamers are made by swapping O and N atom in side chain to switch proton donor and acceptor. Conformer pairs with heavy atom clashes are optimized with 2 cycles of heavy atom relaxation. Approximately 24 to 42 million microstates for generated each MD snapshot input structure.

MCCE allows the protonation state of all groups with the exceptions of those listed in Table S1 to come to equilibrium at pH 7. All 4 Propionic groups are free to titrate. The MCCE analysis of snapshots show: in the $EP_{a3}^-$ trajectory $PRD_a$ is 5% protonated; in the $E^-P_{a3}$ trajectory the $PRA_{a3}$ and $PRD_{a3}$ are 20% and 32% protonated; in the $EP_{a3}$ trajectory the $PRA_{a3}$ and $PRD_{a3}$ are 15% and 20% protonated. All other propionic acids are >99.9% ionized. The crystal structures 1M56 and 2GSM, have all propionic groups fully ionized. Other residues with protonation states in the MCCE equilibrium ensemble that differ from what would be found in solution at pH 7 are given in Table S7.

## S3. Additional characterization inter-cluster connections in CcO hydrogen bond network

The main focus of the paper is to characterize the PLS and its connections to the P-side surface of CcO. The analysis also provides additional information regarding other inter-cluster connections in the CcO hydrogen bond networks, which are described here.

**Glu286 to D channel:** Glu286 is a hub in the CcO network, but its connections depend on local hydration of the Glu286 cavity. Glu286 is always connected to the D-channel with the exception of two snapshots, which have water-mediated contacts just beyond our hydrogen bonding cutoff. With a wet cavity (>5 waters) the Glu side chain generally points upward so connections to the D-channel require more than one water (Table S4). With a dry cavity, Glu286 tends to point down, requiring only 1 to 2 waters to connect to the D channel. Analysis of all 500 frames of the 50 ns $E^{-}P_{a3}$ trajectory shows the Glu286 orientation depends weakly on the cavity hydration with a correlation coefficient $R^2$ of 0.26. The orientation varies more at lower hydration levels, while a wet cavity strongly favors the upward orientation.

**Glu286 to PLS and BNC**: The hydration of the Glu286 cavity also affects connections to the PLS and BNC. The connections from Glu286 to the BNC terminates at the ferryl oxygen of heme $a_3$ or the $Cu_B$ hydroxide, which are the intermediates in the production of water. Connections from Glu286 to the BNC are favored with hydration ≥ 5 waters. Overall, when Glu286 points up, facing the PLS with a well hydrated cavity it can be connected to both PLS and BNC. On the other hand, when Glu286 points down and the cavity is dry and compact, the connection to the both BNC and PLS are broken.

**BNC to PLS:** The connections from PLS to BNC substrate oxygens on heme $a_3$ and $Cu_B$ are puzzling since they can facilitate back-transfer of protons from the PLS to BNC. They are mediated by about 4 to 5 waters (Table S4). Interestingly, the same individual waters are on pathways that connect Glu286 to the BNC and the BNC to PLS. Given the potential connectivity, back proton transfers from PLS to BNC will need to be suppressed by energetic barriers that ultimately control the rates (20). The current study is limited in scope to identifying the pathways. Future work will be needed to address the free energy of barriers to proton leakage.

## Tables

**Table S1.** Protonation and redox states of residues fixed in MD trajectories and MCCE simulations of $EP_{a3}^{-}$, $E^{-}P_{a3}$ and $EP_{a3}$ substates.

| **Substate** | $EP_{a3}^{-}$ | | $E^{-}P_{a3}$ | | $EP_{a3}$ | |
|---|---|---|---|---|---|---|
| **Method** | MD | MCCE | MD | MCCE | MD | MCCE |
| **PRA of heme *a*** | deprotonated | free | deprotonated | free | deprotonated | free |
| **PRD of heme *a*** | deprotonated | free | deprotonated | free | deprotonated | free |
| **PRA of heme $a_3$** | deprotonated | free | protonated | free | protonated | free |
| **PRD of heme $a_3$** | deprotonated | free | deprotonated | free | deprotonated | free |
| **E286** | protonated | protonated | deprotonated | deprotonated | protonated | protonated |
| **heme *a*** | oxidized | | | | | |
| **$Cu_A$** | oxidized | | | | | |
| **heme $a_3$** | $Fe(IV)=O^{2-}$ | | | | | |
| **$Cu_B$** | $Cu(II)\text{-}OH^{-}$ | | | | | |
| **TYR288** | deprotonated ($TyrO^{-}$) | | | | | |
| **Mg** | $Mg^{2+}$ | | | | | |
| **Ca** | $Ca^{2+}$ | | | | | |

MCCE calculations of the X-ray structures use the protonation assignments for the $E^{-}P_{a3}$ substate.

**Table S2.** Numbers of buried waters in input structure and retained after MC sampling.

| Snapshot | Time (ns)[a] | water retained | | | | | |
|---|---|---|---|---|---|---|---|
| | | $E^{-}P_{a3}$ | | $EP_{a3}^{-}$ | | $EP_{a3}$ | |
| | | num_snap[b] | num_MC[c] | num_snap | num_MC | num_snap | num_MC |
| **1** | 0 | 178 | 118.39 | 153 | 98.73 | 167 | 107.09 |
| **2** | 3.5 | 185 | 121.43 | 153 | 101.38 | 180 | 119.15 |
| **3** | 7 | 172 | 113.89 | 151 | 101.36 | 164 | 104.48 |
| **4** | 10.5 | 197 | 127.49 | 155 | 93.46 | 165 | 106.74 |
| **5** | 14 | 166 | 108.63 | 162 | 108.51 | 173 | 108.69 |
| **6** | 20 | 172 | 112.66 | 177 | 110.23 | 172 | 113.63 |
| **7** | 26 | 168 | 110.78 | 176 | 120.60 | 181 | 113.32 |
| **8** | 32 | 190 | 116.26 | 170 | 114.54 | 179 | 117.84 |
| **9** | 38 | 174 | 112.08 | 178 | 114.37 | 169 | 102.42 |
| **10** | 44 | 177 | 115.10 | 176 | 114.51 | 160 | 98.55 |
| **11** | 50 | 180 | 115.41 | 174 | 111.49 | 147 | 96.65 |
| **Average** | - | 178.09 | 115.65 | 165.91 | 108.11 | 168.82 | 108.05 |
| **Stdev** | - | 9.44 | 5.29 | 11.16 | 8.29 | 10.00 | 7.39 |
| | | | | | | | |
| **1M56[d]** | - | 65 | 31.69 | - | - | - | - |
| **2GSM[d]** | - | 119 | 87.95 | - | - | - | - |

[a] Timepoint in the trajectory at which snapshot was taken.

[b] Number of waters retained as input for MCCE calculation after surface waters (with accessible surface area larger than 5%) are removed from MD snapshots.

[c] Average number of waters retained after Grand Canonical MC sampling in MCCE.

[d] Analysis of internal waters in the crystal structures.

**Table S3.** Hydrogen bod donor/acceptor definitions used in this study.

| Donor | Acceptor | Donor and Acceptor |
|---|---|---|
| | | HOH(neutral) |
| ARG (ionized) | | ARG (neutral) |
| HIS (ionized) | | HIS (neutral) |
| LYS (ionized) | | LYS (neutral) |
| | ASP (ionized) | ASP (neutral) |
| | GLU (ionized) | GLU (neutral) |
| | TYR (ionized) | TYR (neutral) |
| | | SER |
| | | THR |
| | | ASN |
| | | GLN |
| TRP | | |
| | PRA (ionized) | PRA (neutral) |
| | PRD (ionized) | PRD (neutral) |
| | TYR288 (ionized) | TRY288(neutral) |
| | heme $a_3$=$O^{2-}$ | $Cu_B$-$OH^-$ |

Residues are defined as hydrogen bond donors and/or acceptors based on their protonation states. heme $a_3$, $Cu_B$ and Tyr288 make up the BiNuclear Center (BNC).

**Table S4** The Glu286 and P-exit cavity hydration and final cluster connectivity for each MD snapshot used for MCCE simulations.

| Snapshot Index | Time(ns)[a] | E286[b] | P-exit[c] | E-CAD distance[d] | PLS→PE | E→PLS | D→E | E→BNC | BNC→PLS |
|---|---|---|---|---|---|---|---|---|---|
| $E^{-}P_{a3}$-8 | 32 | 9 | 7 | 10.9 | T | T | T | F | T |
| $E^{-}P_{a3}$-7 | 26 | 8 | 8 | 10.3 | T | T | T | F | F |
| $E^{-}P_{a3}$-11 | 50 | 8 | 7 | 8.5 | T | T | T | T | T |
| $E^{-}P_{a3}$-1 | 0 | 7 | 8 | 9.9 | T | T | T | T | T |
| $E^{-}P_{a3}$-6 | 20 | 7 | 8 | 9.9 | T | T | T | T | T |
| $E^{-}P_{a3}$-10 | 44 | 7 | 8 | 9.5 | T | T | T | T | T |
| $E^{-}P_{a3}$-2 | 3 | 7 | 7 | 10.5 | T | T | T | T | T |
| $E^{-}P_{a3}$-4 | 10 | 7 | 5 | 9.4 | T | T | T | T | T |
| $E^{-}P_{a3}$-9 | 38 | 6 | 7 | 10.9 | T | T | T | T | T |
| $E^{-}P_{a3}$-5 | 14 | 5 | 7 | 10 | T | T | T | F | F |
| $E^{-}P_{a3}$-3 | 7 | 5 | 5 | 9.9 | T | T | T | T | T |
| $EP_{a3}^{-}$-5 | 14 | 5 | 4 | 9.3 | T | F | T | T | F |
| $EP_{a3}^{-}$-1 | 0 | 5 | 3 | 10.5 | F | F | T | F | F |
| $EP_{a3}^{-}$-4 | 10 | 4 | 7 | 9.5 | T | F | T | F | F |
| $EP_{a3}^{-}$-8 | 32 | 3 | 6 | 10.6 | T | F | T | F | F |
| $EP_{a3}$-7 | 26 | 3 | 6 | 10.9 | T | F | T | F | T |
| $EP_{a3}^{-}$-6 | 20 | 3 | 5 | 10.7 | T | F | T | F | F |
| $EP_{a3}^{-}$-9 | 38 | 2 | 7 | 10.5 | T | F | T | F | F |
| $EP_{a3}^{-}$-11 | 50 | 2 | 7 | 10.8 | T | F | T | F | F |
| $EP_{a3}$-4 | 10 | 2 | 7 | 12.8 | T | F | T | F | T |
| $EP_{a3}$-5 | 14 | 2 | 7 | 12.1 | T | F | T | F | T |
| $EP_{a3}^{-}$-7 | 26 | 2 | 5 | 11.9 | F | F | T | F | F |
| $EP_{a3}$-3 | 7 | 2 | 5 | 12.6 | T | F | F | F | T |
| $EP_{a3}^{-}$-2 | 3 | 2 | 4 | 10.3 | F | F | T | F | F |
| $EP_{a3}^{-}$-3 | 7 | 2 | 3 | 10.1 | F | F | F | T | F |
| $EP_{a3}$-6 | 20 | 1 | 8 | 11 | T | F | T | F | T |
| $EP_{a3}$-1 | 0 | 1 | 7 | 11.5 | T | F | T | F | F |
| $EP_{a3}$-8 | 32 | 1 | 7 | 11.4 | T | F | T | F | T |
| $EP_{a3}$-10 | 44 | 1 | 7 | 12.8 | T | F | T | F | T |
| $EP_{a3}$-9 | 38 | 1 | 6 | 12.1 | T | F | T | F | F |
| $EP_{a3}$-11 | 50 | 1 | 6 | 11 | T | F | T | F | F |
| $EP_{a3}^{-}$-10 | 44 | 1 | 4 | 9.2 | T | F | T | F | F |
| $EP_{a3}$-2 | 3 | 1 | 4 | 11 | T | F | T | F | T |

The table is ordered by the number of waters in Glu286 cavity then by the number of waters in P-exit cavity. The snapshots from the same trajectory have the same font color in first column. Shaded rows correspond to snapshots where Glu286 to PLS connections are always maintained. PLS→PE, E→PLS, D→E, E→BNC, BNC→PLS note the inter-cluster connections between the clusters, with T indicating there are connections and **F** means there are not. E is Glu286. D is D channel and PE is the P-exit cluster.

[a]The time point in nanoseconds of the MD trajectory for each index.

[b]The number of waters within 4.5 Å of either terminal oxygen of Glu286.

[c]The number of unique waters within 3.5 Å of the oxygen and nitrogen termini of Trp95, Asn96 or Thr100.
[d]The Glu286 orientation is defined by the distance in nm between Glu286 CD and heme $a_3$ CAD atoms in Å. Value $\leqslant$ 10.0 Å indicates the Glu286 points upwards towards the PLS. Larger values move it downwards into the D channel.

**Table S5.** The conservation of residues in the loop spanning residues number 167 to 179.

| Res # *Rb.sphaeroides* | Amino Acid Freq. | HSSP Weight |
|---|---|---|
| G167 | G 90% | 0.81 |
| S168 | A 74%, S 5% | 0.6 |
| G169 | G 73% | 0.46 |
| I170 | T 85%, I 0% | 0.71 |
| G171 | G 100% | 1 |
| W172 | W 100% | 1 |
| V173 | T 82%, V 3% | 0.55 |
| L174 | V 63%, L 5% | 0.39 |
| Y175 | Y 99% | 0.99 |
| P176 | P 82% | 0.66 |
| P177 | P 82% | 0.97 |
| L178 | L 97% | 0.95 |
| S179 | S 63% | 0.6 |

The sequences of the loop region are aligned using the 4110 sequences in the HSSP database for 1M56 (Homology Derived Secondary Structure of Proteins version 2.0 2011) [22]. The residue ID follows *Rb. sphaeroides* CcO sequence numbering. The amino acid profile shows the most frequent amino acid type observed in this position, following by its percentage. Rows highlighted in grey correspond to amino acid positions where the predominant amino acid type is different from that in *Rb. sphaeroides*. The sequence positions 171, 172, 175, 177, 178 (GWYPV) in the Trp loop are highly conserved.

**Table S6.** The specific residues making the inter-cluster connections via 4 or fewer waters in MCCE analysis of snapshots in the highly connected $E^{-}P_{a3}$ trajectory (Fig 2B).

| N-surface---D-channel connections | | | |
|---|---|---|---|
| **path#** | **donor** | **acceptor** | **% occurrence** |
| | **N-surface** | **D-channel** | |
| **1** | H26 | D132 | 100 |
| **2** | H549 | | 36 |
| **3** | N25 | | 45 |
| **4** | T550 | | 27 |

| D-channel---BNC connections | | | |
|---|---|---|---|
| **path#** | **donor** | **acceptor** | **% occurrence** |
| | **BNC** | **D-channel** | |
| **1** | $Cu_B$ | E286 | 73 |

| D-channel---PLS connections | | | |
|---|---|---|---|
| **path#** | **donor** | **acceptor** | **% occurrence** |
| | **PLS** | **D-channel** | |
| **1** | D412 | E286 | 18 |
| **2** | E254 | | 27 |
| **3** | $PRA_{a3}$ | | 27 |
| **4** | $PRD_{a3}$ | | 36 |
| **5** | R481 | | 100 |
| **6** | R482 | | 73 |
| **7** | W172 | | 9 |
| **8** | Y175 | | 27 |

| BNC---PLS connections | | | |
|---|---|---|---|
| **path#** | **donor** | **acceptor** | **% occurrence** |
| | **BNC** | **PLS** | |
| **1** | $Cu_B$ | D412 | 9 |
| **2** | | $PRA_{a3}$ | 9 |
| **3** | | $PRD_{a3}$ | 73 |
| | **PLS** | **BNC** | |
| **4** | D412 | $Cu_B$ | 9 |
| **5** | $PRD_{a3}$ | Heme $a_3$ | 9 |
| **6** | | $Cu_B$ | 9 |
| **7** | R481 | Heme $a_3$ | 36 |
| **8** | | $Cu_B$ | 64 |
| **9** | R482 | Heme $a_3$ | 27 |
| **10** | | $Cu_B$ | 45 |

| PLS---PLS' connections | | | |
|---|---|---|---|
| path# | donor | acceptor | % occurrence |
| | PLS | PLS' | |
| 1 | R481 | $PRA_a$ | 100 |
| | PLS' | PLS | |
| 2 | R52 | $PRD_a$ | 9 |
| 3 | Y414 | | 9 |

| PLS---P-exit connections | | | |
|---|---|---|---|
| path# | donor | acceptor | % occurrence |
| | PLS | P-exit | |
| 1 | R481 | T100 | 9 |
| 2 | R482 | N96 | 64 |
| 3 | | T100 | 73 |
| 4 | | S168 | 9 |
| 5 | | S186 | 9 |
| 6 | W172 | N96 | 73 |
| 7 | | T100 | 82 |
| 8 | | S168 | 9 |
| 9 | | S186 | 9 |
| | P-exit | PLS | |
| 10 | N96 | R482 | 9 |
| 11 | | $PRD_a$ | 27 |
| 12 | S168 | $PRD_a$ | 9 |
| 13 | S186 | $PRD_a$ | 18 |
| 14 | T100 | R482 | 18 |
| 15 | | E254B | 9 |
| 16 | | $PRD_a$ | 73 |
| 17 | W95 | R482 | 9 |
| 18 | | $PRD_a$ | 91 |

| P-exit---P-exit' | | | |
|---|---|---|---|
| path# | donor | acceptor | % occurrence |
| | P-exit | P-exit' | |
| 1 | H93 | N91 | 36 |
| 2 | | Q165 | 36 |
| 3 | R257 | D188 | 100 |
| 4 | | N258 | 100 |
| 5 | Y262B | D485 | 73 |
| 6 | | E488 | 9 |
| | P-exit' | P-exit | |
| 8 | D485 | N170B | 9 |
| 9 | | Y262B | 9 |
| 10 | E488 | Y262B | 9 |
| 11 | N91 | H93 | 18 |
| 13 | Q165 | H93 | 45 |

| PLS---Cluster 1 connections | | | |
|---|---|---|---|
| **path#** | **donor** | **acceptor** | **% occurrence** |
| | **PLS** | **Cluster 1** | |
| **1** | S218B | S253B | 9 |
| | **Cluster 1** | **PLS** | |
| **2** | S253B | S218B | 9 |

| Cluster 1---Cluster 2 connections | | | |
|---|---|---|---|
| **path#** | **donor** | **acceptor** | **% occurrence** |
| | **Cluster 2** | **Cluster 1** | |
| **1** | Q126B | Y409 | 9 |
| **2** | | T220B | 9 |
| **3** | | K227B | 9 |

The hydrogen bond networks are characterized highly interconnected clusters (residues given in Table 1), with sparse connections between them. The residues that anchor these connections in snapshots from the $E^{-}P_{a3}$ state network with four or fewer waters are listed here. As the $E^{-}P_{a3}$ state has well hydrated Glu286 and P-exit cavities, inter-cluster connections are found between all clusters. In addition, the inter-cluster connections of the two weakly connected clusters (Cluster 1 and Cluster 2) are provided. The occupancy shows the percentage of the 11 snapshots that are connected through the 2 residues. Connections are noted without consideration of the direction of the hydrogen bonding linking the clusters.

**Table S7.** The residues with MCCE equilibrated charge states that differ from those found for the residue in solution at pH 7.

| **MD snapshots (of 33 total snapshots)** | | |
|---|---|---|
| **RESIDUE** | # models | AVE_CRG[a] |
| **ASP407** | 33 | 0.00 |
| **HIS534** | 33 | 0.81 |
| **HIS549** | 33 | 0.49 |
| **HIS67** | 33 | 0.89 |
| **HIS96B** | 33 | 0.31 |
| **LYS362** | 33 | 0.00 |
| **LYS442** | 33 | 0.06 |
| **HIS93** | 32 | 0.12 |
| **GLU54** | 21 | -0.05 |
| **ASP256** | 15 | -0.82 |
| **GLU254B** | 13 | 0.00 |
| **LYS227B** | 11 | 0.68 |
| **Crystal structures (1M56 and 2GSM)** | | |
| **ASP407** | 2 | 0.00 |
| **ASP412** | 2 | 0.00 |
| **HIS534** | 2 | 0.69 |
| **HIS549** | 2 | 0.85 |
| **HIS93** | 2 | 0.55 |
| **HIS96B** | 2 | 0.49 |
| **LYS362** | 2 | 0.00 |
| **LYS454** | 2 | 0.90 |
| **LYS227B** | 2 | 0.08 |

[a] Average MCCE equilibrated charge of residue in snapshots or crystal structures.

**Figure S1. Hydrogen bond networks as a function of the number of waters allowed to mediate hydrogen bonds between residues**. Analysis is based on 11 snapshots from $E^{-}P_{a3}$ trajectory. The same input data is used for each network. Connections with zero waters show side chains in direct contact. The position and color of the residue (nodes) are the same as in Fig2B.

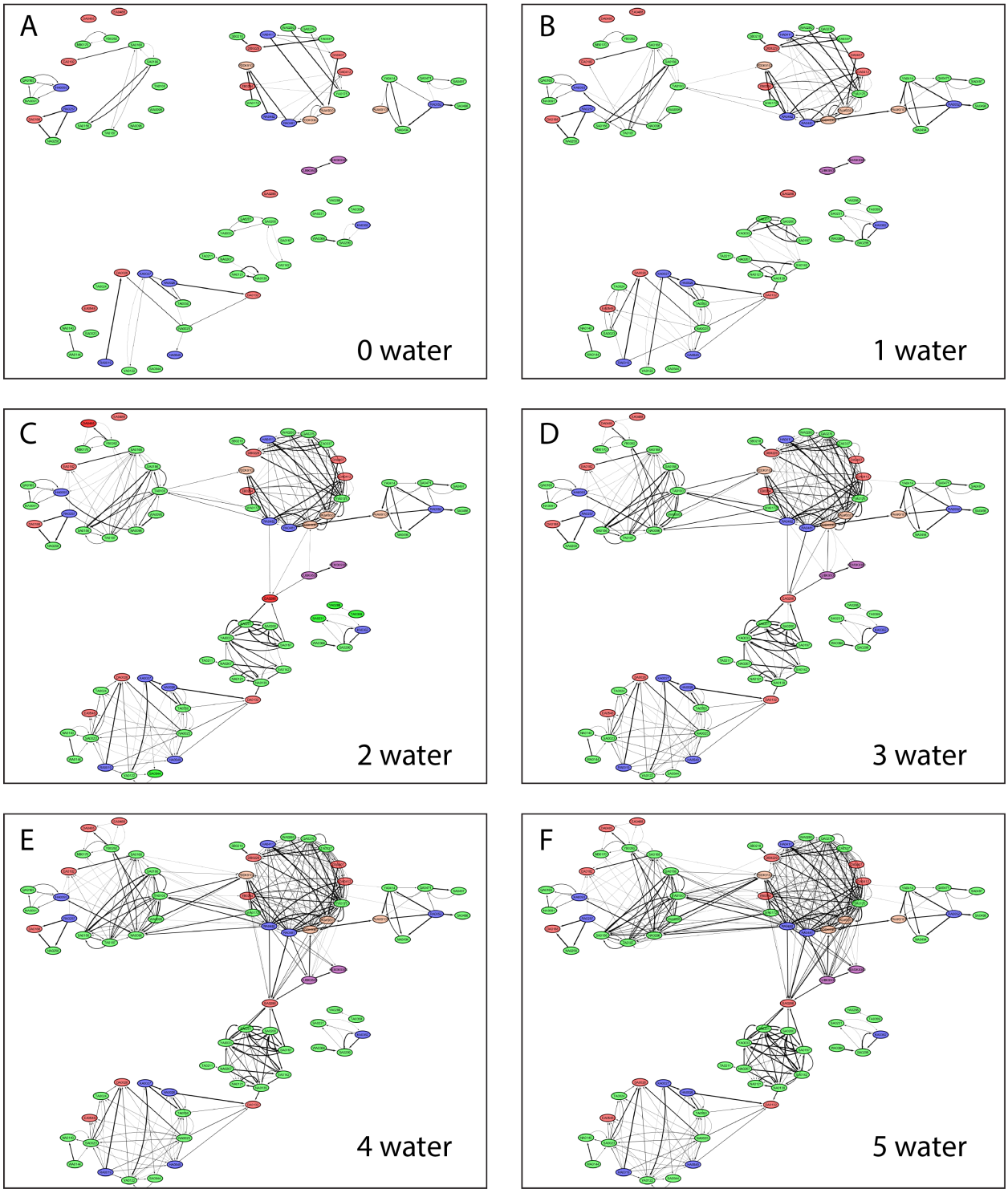
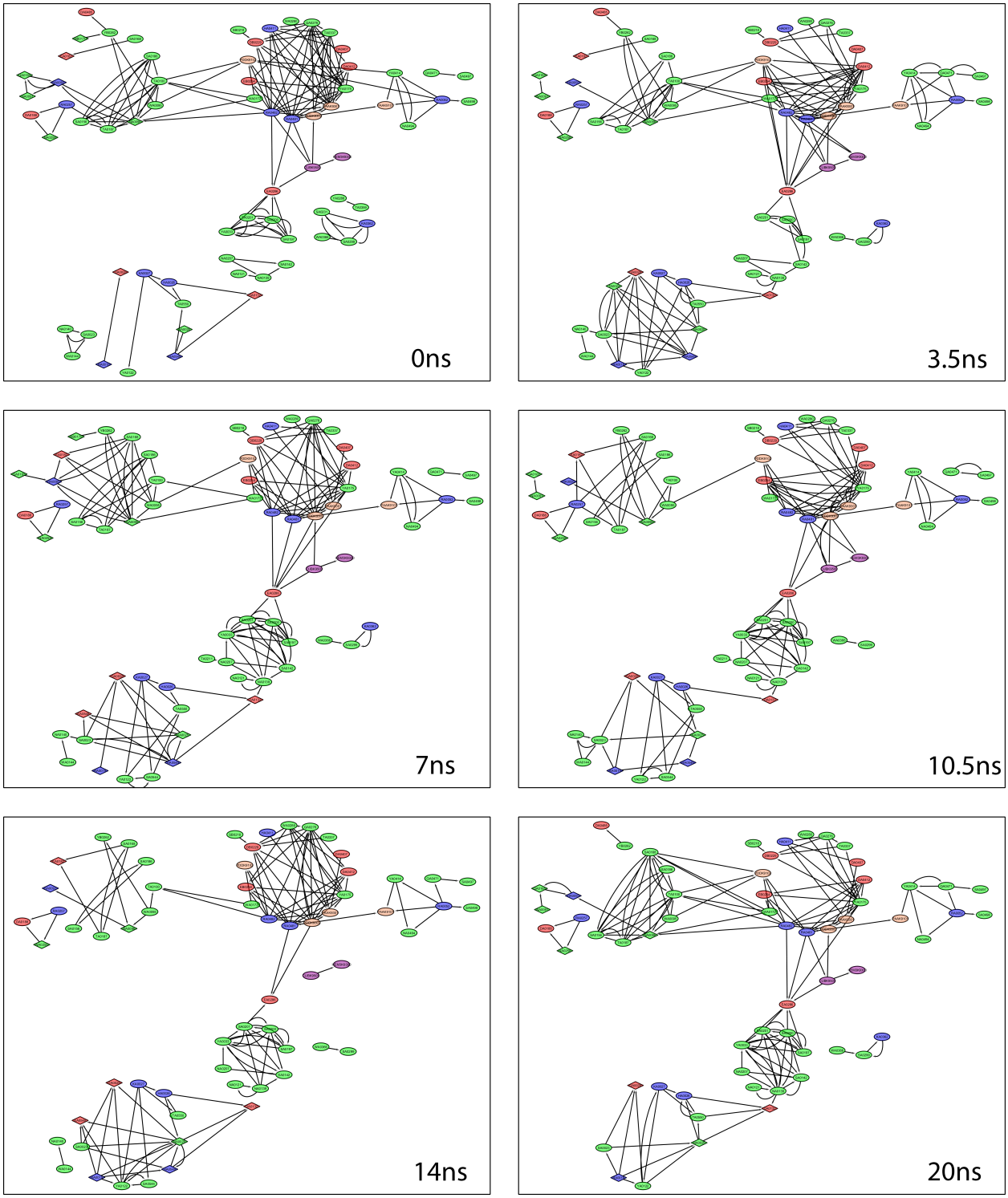
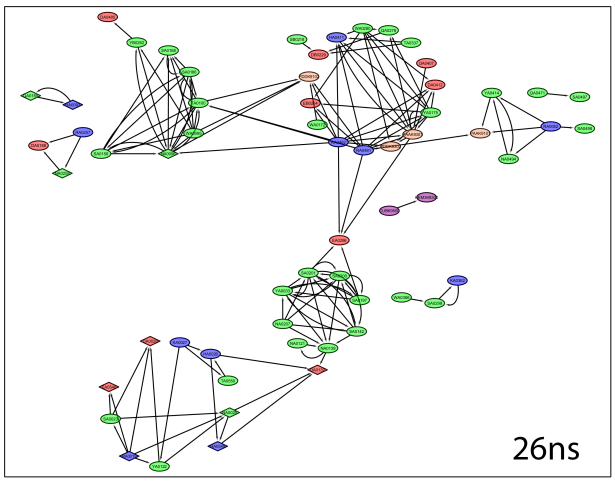
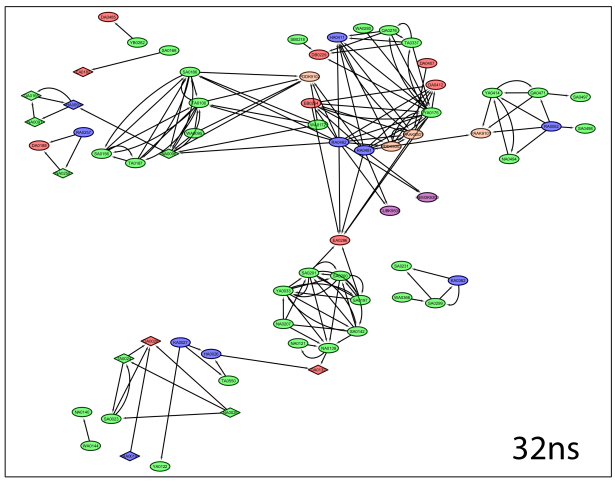
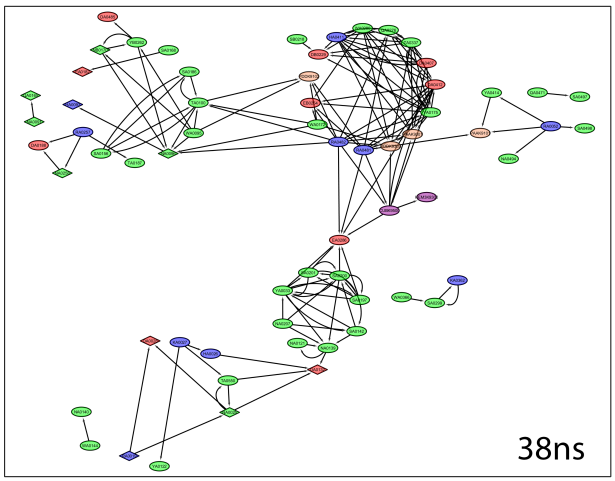


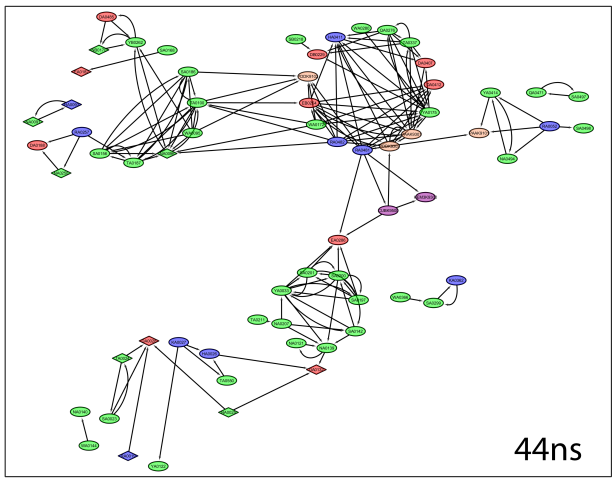
**Figure S2. Networks obtained from individual snapshots from the $E^{-}P_{a3}$ trajectory.** Each connection represents a direct hydrogen bond between residues or an indirect connection mediated by up to 4 waters. The position and color of the residue (nodes) are the same as in Fig2B and S1. Trajectory time is indicated in each figures.




26ns


32ns


38ns


44ns

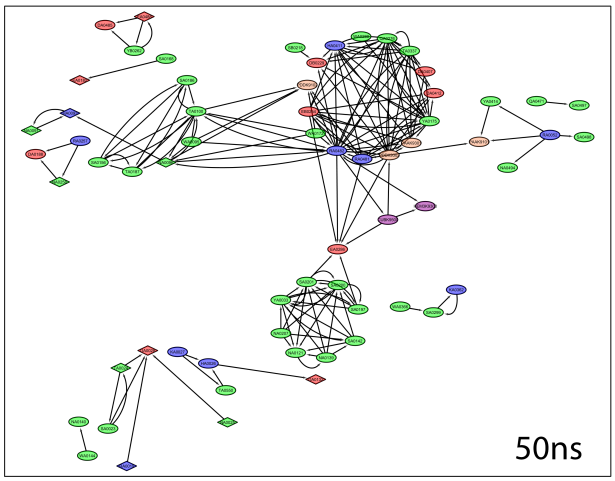

50ns

***Figure S3. Hydrogen bond network of crystal structure****. (A) PDB:1M56 and (B) PDB:2GSM. MCCE calculations are performed using $E^{-}P_{a3}$ state (Table S1) as this provided the most connected state in simulations. The position and color of the residue (nodes) are the same as in Fig2B and S2, S3.*

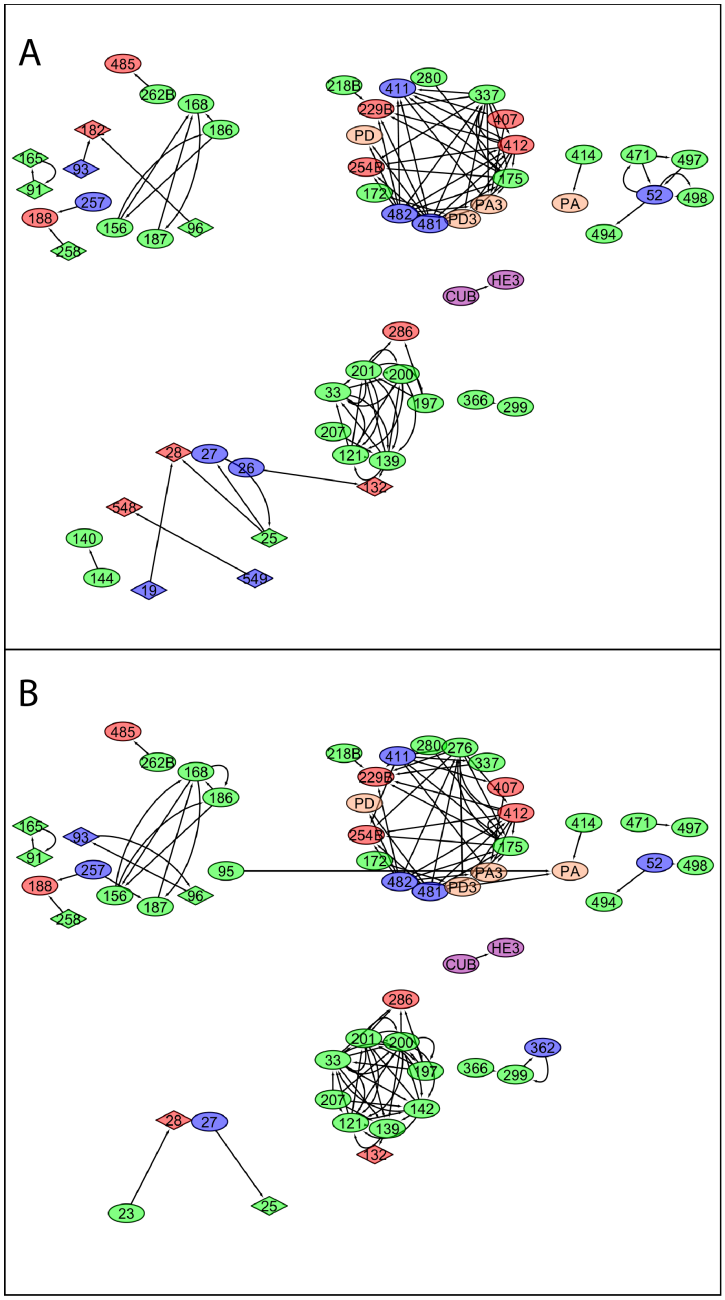


**Figure S4. Key nodes in hydrogen bond network (with four-water connection) in the different substates**. Here the node sizes is scaled by betweenness centrality and edge width by edge betweenness[23]. The betweenness centrality for each node is the sum of the fraction of all-pairs of shortest paths that pass through the node. The edge betweenness centrality is the sum of the fraction of all-pairs shortest paths that pass through the edge. Nodes with more control over the network have higher betweenness centrality and edge betweenness. The networks are shown for MCCE network analysis of MD trajectories in different substates of CcO. $E^{-}P_{a3}$ (A) has well hydrated Glu286 and P-exit cavities. In the $EP_{a3}$ (C) snapshots the Glu286 cavity is dry, while $EP_{a3}^{-}$ (B) fluctuates between dry and partially hydrated cavities.

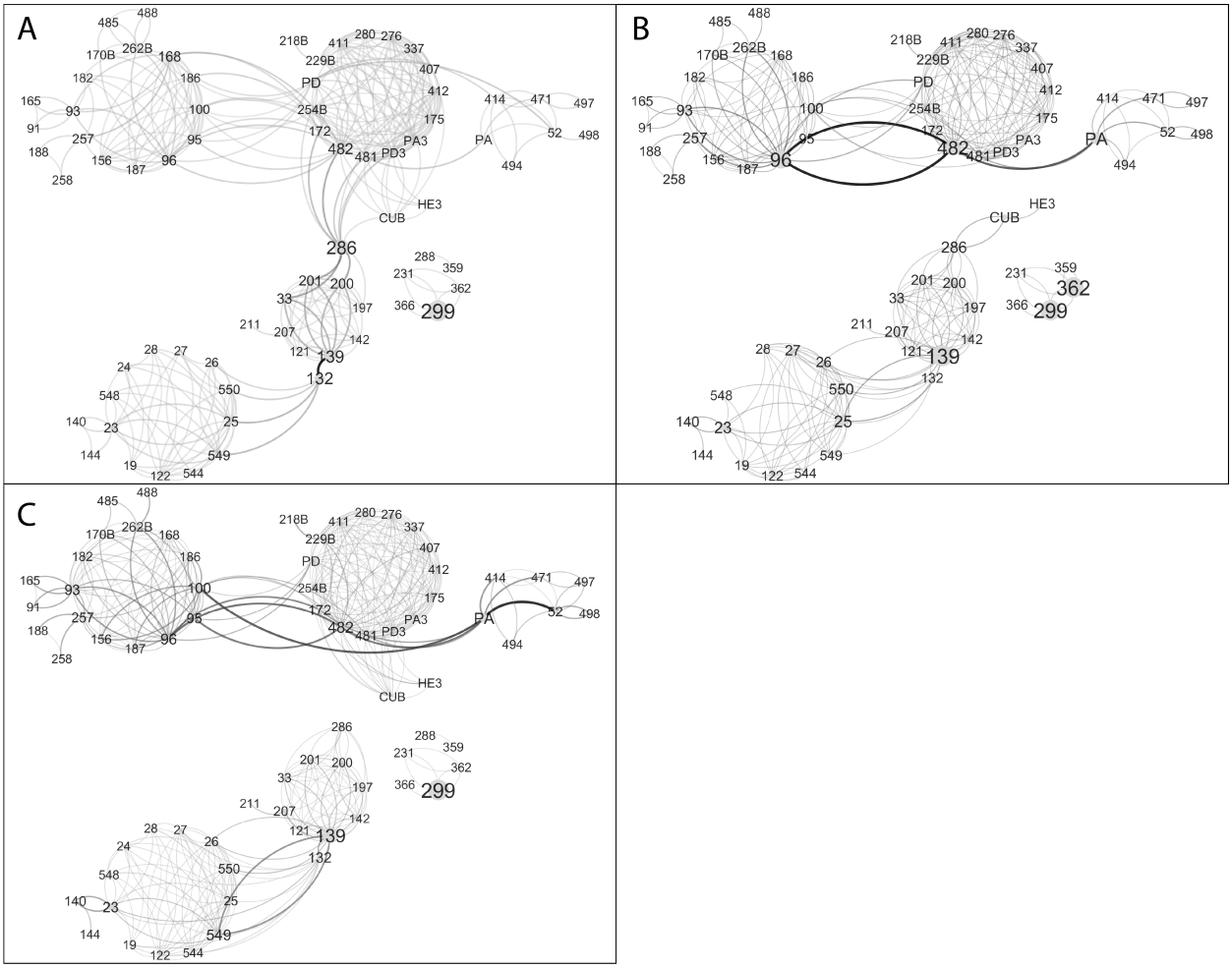


**Figure S5. Proton exit pathways found in network analysis extending from $PRA_a$ to the surface residues.** Via (A) His93 and Glu182 or (B) the nearby Tyr262B obtained from the analysis of continuously hydrogen bonded residues (in the hydrogen bond network) found in individual MCCE microstates. These networks extend over approximately (A) 19 Å and (B) 30 Å.

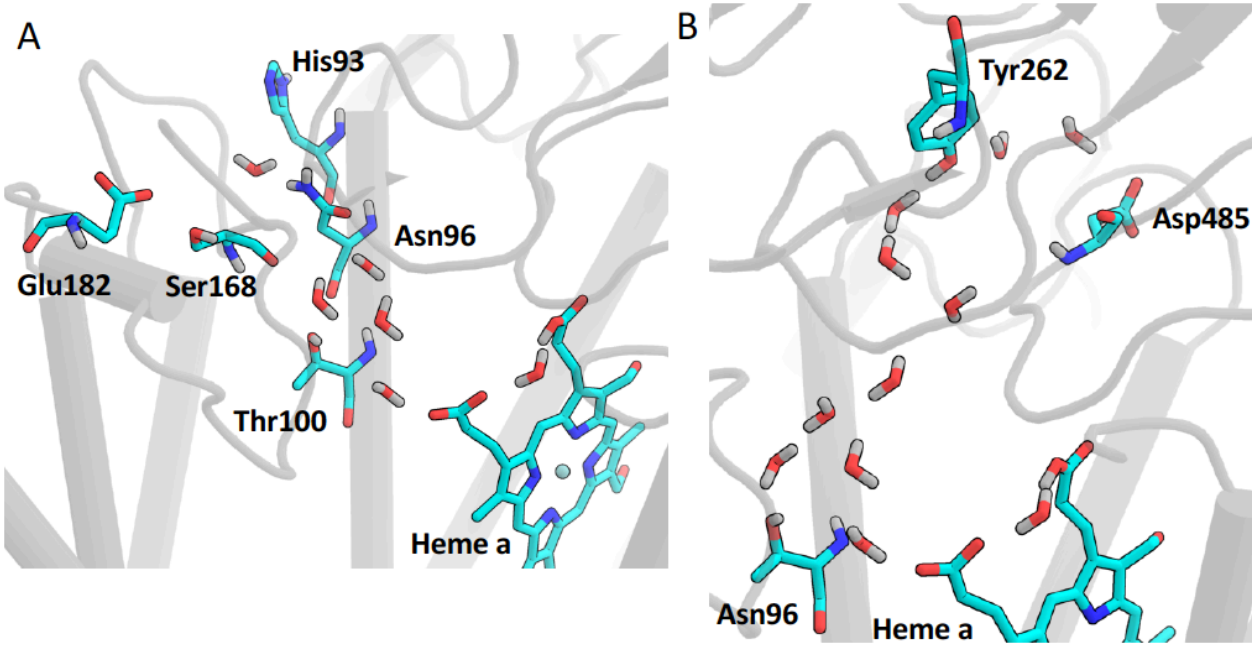
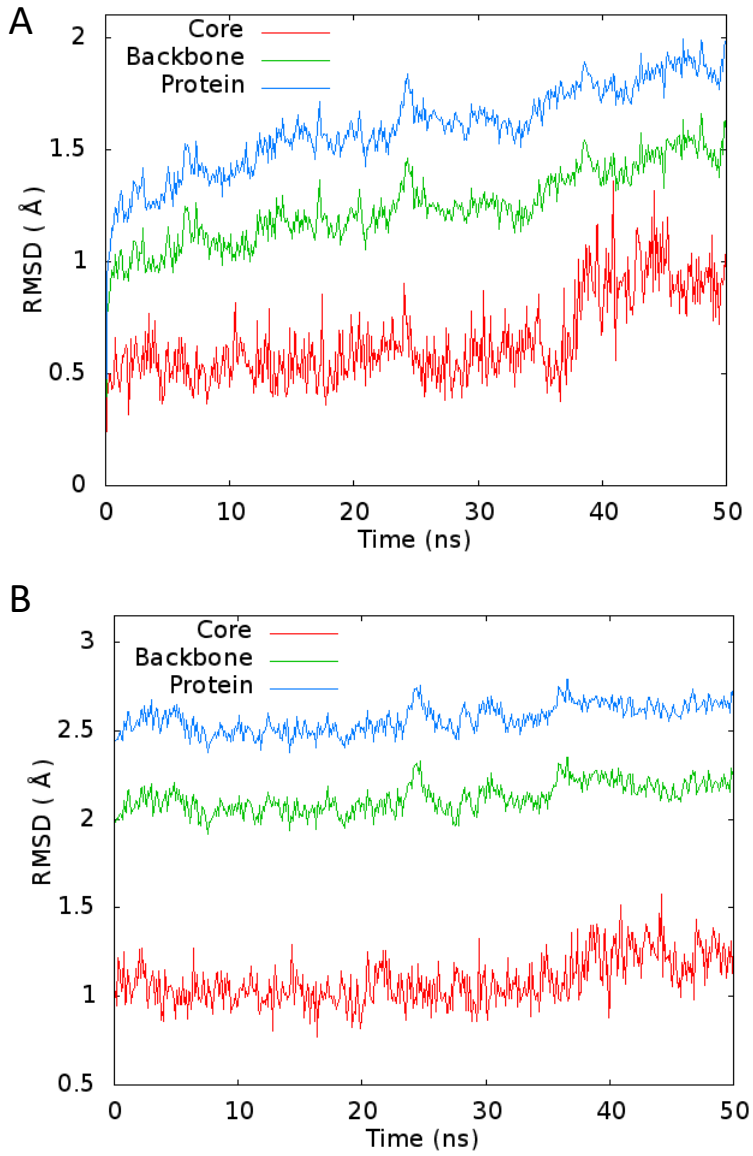


**Figure S6. RMSD of the protein during the 50ns MD trajectory in $EP_{a3}$ state.**. Calculations relative to (A) initial $EP_{a3}^{-}$ structure and (B) the crystal structure (PDB: 1M56). Core: heme a, BNC and the backbone of Glu286; backbone: all backbone atoms (C H O N); protein: the entire protein including all the side chains.



**References**


[1] M. Svensson-Ek, J. Abramson, G. Larsson, S. Törnroth, P. Brzezinski, S. Iwata, The X-ray crystal structures of wild-type and EQ(I-286) mutant cytochrome c oxidases from *Rhodobacter sphaeroides*, J. Mol. Biol. 321 (2002) 329–339.

[2] Y. Song, E. Michonova-Alexova, M.R. Gunner, Calculated proton uptake on anaerobic reduction of cytochrome C oxidase: is the reaction electroneutral?, Biochemistry. 45 (2006) 7959–7975. doi:10.1021/bi052183d.

[3] P. Goyal, J. Lu, S. Yang, M.R. Gunner, Q. Cui, Changing hydration level in an internal cavity modulates the proton affinity of a key glutamate in cytochrome c oxidase, Proc. Natl. Acad. Sci. U.S.A. 110 (2013) 18886–18891. doi:10.1073/pnas.1313908110.

[4] J. Lu, M.R. Gunner, Characterizing the proton loading site in cytochrome c oxidase, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) 12414–12419. doi:10.1073/pnas.1407187111.

[5] N. Ghosh, X. Prat-Resina, M.R. Gunner, Q. Cui, Microscopic pKa analysis of Glu286 in cytochrome c oxidase (*Rhodobacter sphaeroides*): toward a calibrated molecular model, Biochemistry. 48 (2009) 2468–2485. doi:10.1021/bi8021284.

[6] A.D. MacKerell, D. Bashford, M. Bellott, R.L. Dunbrack, J.D. Evanseck, M.J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F.T. Lau, C. Mattos, S. Michnick, T. Ngo, D.T. Nguyen, B. Prodhom, W.E. Reiher, B. Roux, M. Schlenkrich, J.C. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiórkiewicz-Kuczera, D. Yin, M. Karplus, All-atom empirical potential for molecular modeling and dynamics studies of proteins, J Phys Chem B. 102 (1998) 3586–3616. doi:10.1021/jp973084f.

[7] A.D. Mackerell, M. Feig, C.L. Brooks, Extending the treatment of backbone energetics in protein force fields: limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, J Comput Chem. 25 (2004) 1400–1415. doi:10.1002/jcc.20065.

[8] J.B. Klauda, R.M. Venable, J.A. Freites, J.W. O'Connor, D.J. Tobias, C. Mondragon-Ramirez, I. Vorobyov, A.D. MacKerell, R.W. Pastor, Update of the CHARMM all-atom additive force field for lipids: Validation on six lipid types, J Phys Chem B. 114 (2010) 7830–7843. doi:10.1021/jp101759q.

[9] M.P. Johansson, V.R.I. Kaila, L. Laakkonen, Charge parameterization of the metal centers in cytochrome c oxidase, J. Comput. Chem. 29 (2008) 753–767. doi:10.1002/jcc.20835.

[10] S. Pronk, S. Páll, R. Schulz, P. Larsson, P. Bjelkmar, R. Apostolov, M.R. Shirts, J.C. Smith, P.M. Kasson, D. van der Spoel, B. Hess, E. Lindahl, GROMACS 4.5: a high-throughput and highly parallel open source molecular simulation toolkit, Bioinformatics. 29 (2013) 845–854. doi:10.1093/bioinformatics/btt055.

[11] T. Darden, D. York, L. Pedersen, Particle mesh Ewald: An N·log(N) method for Ewald sums in large systems, The Journal of Chemical Physics. 98 (1993) 10089–10092. doi:10.1063/1.464397.

[12] B. Hess, H. Bekker, H.J.C. Berendsen, J.G.E.M. Fraaije, LINCS: A linear constraint solver for molecular simulations, J. Comput. Chem. 18 (1997) 1463–1472. doi:10.1002/(SICI)1096-987X(199709)18:12<1463::AID-JCC4>3.0.CO;2-H.

[13] S. Nosé, A molecular dynamics method for simulations in the canonical ensemble, Molecular Physics. 52 (1984) 255–268. doi:10.1080/00268978400101201.

[14] W.G. Hoover, Canonical dynamics: Equilibrium phase-space distributions, Phys. Rev. A. 31 (1985) 1695–1697. doi:10.1103/PhysRevA.31.1695.

[15] M. Parrinello, A. Rahman, Polymorphic transitions in single crystals: A new molecular dynamics method, Journal of Applied Physics. 52 (1981) 7182–7190. doi:10.1063/1.328693.

[16] L. Qin, C. Hiser, A. Mulichak, R.M. Garavito, S. Ferguson-Miller, Identification of conserved lipid/detergent-binding sites in a high-resolution structure of the membrane protein cytochrome c oxidase, Proc. Natl. Acad. Sci. U.S.A. 103 (2006) 16117–16122. doi:10.1073/pnas.0606149103.
[17] Y. Song, J. Mao, M.R. Gunner, MCCE2: improving protein pKa calculations with extensive side chain rotamer sampling, J Comput Chem. 30 (2009) 2231–2247. doi:10.1002/jcc.21222.
[18] Y. Song, M.R. Gunner, Halorhodopsin pumps Cl- and bacteriorhodopsin pumps protons by a common mechanism that uses conserved electrostatic interactions, Proc. Natl. Acad. Sci. U.S.A. 111 (2014) 16377–16382. doi:10.1073/pnas.1411119111.
[19] Y. Song, J. Mao, M.R. Gunner, Calculation of proton transfers in Bacteriorhodopsin bR and M intermediates, Biochemistry. 42 (2003) 9875–9888. doi:10.1021/bi034482d.
[20] M.R. Gunner, X. Zhu, M.C. Klein, MCCE analysis of the pKas of introduced buried acids and bases in staphylococcal nuclease, Proteins. 79 (2011) 3306–3319. doi:10.1002/prot.23124.
[21] J. Zhang, M.R. Gunner, Multiconformation continuum electrostatics analysis of the effects of a buried Asp introduced near heme a in *Rhodobacter sphaeroides* cytochrome c oxidase, Biochemistry. 49 (2010) 8043–8052. doi:10.1021/bi100663u.
[22] R. Schneider, A. de Daruvar, C. Sander, The HSSP database of protein structure-sequence alignments., Nucleic Acids Res. 25 (1997) 226–230.
[23] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, T. Ideker, Cytoscape: a software environment for integrated models of biomolecular interaction networks, Genome Res. 13 (2003) 2498–2504. doi:10.1101/gr.1239303.